

Towards model selection for local log-density estimation. Fisher and Wilks-type theorems.*

Sergey Dovgal^{1,2}

¹ Moscow Institute of Physics and Technology,

9 Institutskiy per., Dolgoprudny, Moscow Region, 141700, Russian Federation;

² Institute for Information Transmission Problems (RAS),

Bolshoy Karetny per. 19, buld.1, Moscow 127051, Russian Federation;

vit.north-at-gmail.com, dovgal-at-phystech.edu

Abstract. The aim of this research is to make a step towards providing a tool for model selection for log-density estimation. The author revisits the procedure for local log-density estimation suggested by Clive Loader (1996) and extends the theoretical results to finite-sample framework with the help of machinery of Spokoiny (2012). The results include bias expression from “deterministic” counterpart and Fisher and Wilks-type theorems from “stochastic”. We elaborate on bandwidth trade-off $h(n) = \arg \min O(h^p) + O_p(1/\sqrt{nh^d})$ with explicit constants at big O notation. Explicit expressions involve (i) true density function and (ii) model that is selected (dimension, bandwidth, kernel and basis, e.g. polynomial). Existing asymptotic properties directly follow from our results. From the expressions obtained it is possible to control “the curse of dimension” both from the side of log-density smoothness and the inner space dimension.

1 Introduction

There is a famous trade-off between the parameters of the model: bandwidth, polynomial degree, the basis set, the kernel function. In the *linear kernel density estimation* procedure (Parzen–Rozenblatt) [4], the choice of the kernel function is very important for asymptotic rates. For example, if one introduces a *risk* at point x_0 for given density estimator, then one can state the existence of *minimax estimator*, which requires some special kernels (for example based on Legendre polynomials for quadratic risk) and particular dependence $h = h(n)$ in order to minimize the risk.

Loader’s procedure which we consider, has its advantages and disadvantages. Its main disadvantage is its computational complexity: in order to compute the estimate, we need to implement a convex optimization procedure, where each step requires numerical computing of some multidimensional integral. However,

* This work was done during the Master program at MIPT under supervision of Vladimir Spokoiny. I am very grateful to him for his support during my university studies.

they have implemented a `locfit` R-package, and we refer to [3] for their experimental results. Advantage of the procedure is that regardless of the kernel function, this estimator always provides the minimax optimal rates, the same as for respective linear estimators with special kernels (they are referred to as *kernels of order p* in [4]). Its second advantage is that (in case of polynomial basis) we estimate the derivatives of log-density in addition to the value of log-density.

Another reason for developing finite-sample bounds for this particular estimator, is the use of the *quasi-likelihood* concept: we were able to apply ideas of Spokoiny [1] for finite-sample estimation.

This study allowed to choose the “best kernel” according to our finite-sample bounds. In the case of pointwise estimation, the answer is probably the *indicator kernel*, but it is still unclear, whether it is the same for uniform bounds for multi-point density estimation. The best bandwidth can be chosen by the familiar expression

$$h(n) = \arg \min_h \left(O(h^p) + O_p((nh^d)^{-1/2}) \right) , \quad (1.1)$$

where p stands for smoothness and d for dimension, n is a sample size. Since we provide explicit constants, it becomes possible to choose this minimum explicitly.

We also point out that despite the work that has been done, it is still not enough to provide data-driven procedure for construction of confidence intervals or confidence bands. Usually Fisher and Wilks theorems are used to construct confidence interval at point or a confidence band at some region $x \in I$, but in order to choose the correct data-driven quantile function, *bootstrap* provides a substantial (asymptotic and non-asymptotic) refinement in comparison with more conservative tools. Hopefully, the future research will give answers to these questions.

1.1 Key objects and estimation procedure

Loader [3] considers the following idea. Suppose the data $X_1, X_2, \dots, X_n \in \mathbb{R}^d$ is observed, where X_i are i.i.d. drawn from density function f . Our goal is to construct the estimate $\hat{f}(x_0)$ of the unknown density at given point x_0 . The ordinary likelihood for density function is defined by equation $L(f) = \prod_{i=1}^n f(X_i)$. We

restrict ourselves to density functions that satisfy $\int_{\mathbb{R}} f(x)dx = 1$. However, maximum for this likelihood over functions f , is attained at sum of delta-functions — this is the reason why we impose further smoothness restrictions.

Note that the expectation of likelihood has the nice property of having maximum in true density function f^* :

$$f^* = \arg \max_{f \in L_2} \mathbf{E}L(f) . \quad (1.2)$$

Let us change the procedure in the way so it can become more practical: consider (i) localization $[x_0 - h, x_0 + h]$ with the change of variables $t = \frac{x - x_0}{h}$ and (ii) choose some finite basis $\psi_0(t), \dots, \psi_{p-1}(t)$ for the unknown log-density

function in the interval $t \in [-1, 1]$ (or the cube $[-1, 1]^d$ in case of multidimensional estimation). Let us also introduce *log-likelihood function* parametrized by some vector $\boldsymbol{\theta}$:

$$L(\boldsymbol{\theta}; \mathbf{X}, x_0, h) = \sum_{i=1}^n K_i \boldsymbol{\Psi}_i^\top \boldsymbol{\theta} - n \int K \exp(\boldsymbol{\Psi}^\top \boldsymbol{\theta}) dx , \quad (1.3)$$

where $\boldsymbol{\Psi}_i = (\psi_0(T_i), \psi_1(T_i), \dots, \psi_{p-1}(T_i))^\top$, $\boldsymbol{\theta} = (\theta_0, \theta_1, \dots, \theta_{p-1})$, $K_i = K(T_i)$, $T_i = \frac{X_i - x_0}{h}$, $dx = h^d dt$. The principal example in the current article will be the case of one-dimensional local polynomial estimation with an indicator kernel, where $\psi_k(t) = t^k$, $K(t) = [-1 \leq t \leq 1]$. We also discuss generalizations to the d -dimensional case throughout this article.

The motivation for the functional (1.3) is the following: first terms stands for basis approximation of given density function, and the second terms stands for Lagrange-type penalty.

Then we define $\tilde{\boldsymbol{\theta}}$ — *maximum likelihood estimator*, $\boldsymbol{\theta}^*$ — *target biased parameter*:

$$\tilde{\boldsymbol{\theta}} = \arg \max_{\boldsymbol{\theta}} L(\boldsymbol{\theta}), \quad \boldsymbol{\theta}^* = \arg \max_{\boldsymbol{\theta}} \mathbf{E}L(\boldsymbol{\theta}) , \quad (1.4)$$

We will also define *target unbiased parameter* $\boldsymbol{\theta}^\bullet(h)$ later through *small bias condition*. Since true log-density function isn't necessarily equal to the finite sum of basis functions, in practice one can choose any known approximation as an unbiased parameter. As an illustration, in the case of one-dimensional local polynomial estimation, unbiased parameter can be chosen as first p terms in Taylor expansion of $\log f(x)$ near x_0 :

$$\boldsymbol{\theta}_j^\bullet(h) = \frac{h^j}{j!} \frac{\partial^j \log f(x)}{\partial x^j} \Big|_{x=x_0}, \quad j = 0, 1, \dots, p-1 . \quad (1.5)$$

We will require that the first element of the basis is constant, $\psi_0 \equiv 1$, so that θ_0 usually corresponds to the sought-for log-density: $\theta_0^\bullet = \log f(x_0)$. If the elements of the basis are linearly dependent, then it is not possible to perform the estimation procedure. During the proofs we will use auxilliary parameter defined by

$$\boldsymbol{\theta}^\circ = (\theta_0^\bullet, 0, \dots, 0) = (\log f(x_0), 0, \dots, 0) . \quad (1.6)$$

Below we introduce the objects from finite-sample theory of Spokoiny: the *information matrix* D_n^2 , the *score vector* $\boldsymbol{\xi}$ and the *variance matrix* V_n^2 . The index n stands for sample size, though, in all the statements the value nh^d will be used as an *effective sample size*.

$$\begin{aligned} D_n^2 &= -\nabla^2 \mathbf{E}L(\boldsymbol{\theta}^*) , & D_n^2(\boldsymbol{\theta}) &= -\nabla^2 \mathbf{E}L(\boldsymbol{\theta}) , \\ \boldsymbol{\xi} &= D_n^{-1} \nabla L(\boldsymbol{\theta}^*) , & V_n^2 &= \text{Var}(\nabla L(\boldsymbol{\theta}^*)) . \end{aligned} \quad (1.7)$$

In general case of multidimensional density estimation, $\mathcal{X} \subseteq \mathbb{R}^d$ we denote

$$d_0^2(\boldsymbol{\theta}) = (nh^d)^{-1} D_n^2(\boldsymbol{\theta}) , \quad (1.8)$$

where matrix $d_0^2(\boldsymbol{\theta})$ doesn't depend on n, h .

Since stochastic part of L is linear on $\boldsymbol{\theta}$, the stochastic part of gradient ∇L doesn't depend on the argument: $\nabla L(\boldsymbol{\theta}^*) - \mathbf{E}\nabla L(\boldsymbol{\theta}^*) \equiv \nabla L(\boldsymbol{\theta}) - \mathbf{E}\nabla L(\boldsymbol{\theta})$. We also consider matrix $V_n^2(f)$, which describes the variance under the true measure $f(x)$, depending on various functions f :

$$V_n^2(f) = \text{Var}_f(\nabla L) . \quad (1.9)$$

We also introduce the concentration neighbourhood for $\tilde{\boldsymbol{\theta}}$:

$$\Theta_n(\mathbf{z}) = \left\{ \boldsymbol{\theta} : \|D_n(\boldsymbol{\theta} - \boldsymbol{\theta}^*)\| \leq \mathbf{r}_0(\mathbf{z}) \right\} . \quad (1.10)$$

The concentration radius $\mathbf{r}_0(\mathbf{z})$ will be described below. In particular, the concentration condition (\mathcal{C}) describes upper bounds for \mathbf{r}_0 , and this condition is checked in section 5.

The quantile function for χ^2 -like distributions is given by lemma 12:

$$\zeta(p, \mathbf{z}) = 2\mathfrak{a}\nu_0(\sqrt{p} + \sqrt{2\mathbf{z}}), \quad \mathbf{z} \leq \mathfrak{g}^2/4 , \quad (1.11)$$

where the constants \mathfrak{a}, ν_0 are given by the conditions (\mathcal{I}) and (ED_0) .

1.2 Structure of the article

In order to prove theorems 1, 2 and 3 we need to apply finite-sample machinery of Spokoiny [1], and then check all the conditions. Thus, we state these conditions in section 2 and explain what they mean. This is referred to as “level 1”. When we check these conditions, we want to express them in terms of the model (basis, dimension, bandwidth, kernel and smoothness) and in terms of some unknown true variables: density value at the point x_0 , oscillation and bias. This is referred to as “level 2”, since our logic is clearly separated into layers.

After some preparation in form of conditions and constants we state and prove the main theorems in section 4. These theorems are quite general and are applicable for a wide range of models, see [12, 14]. Some constants are separated from the formulation of the main theorems to keep the presentation more clear. The final expressions can be obtained combining the theorems and results from section 5, where we check the conditions in form of lemmas. The theorem 4 is a separate result, and doesn't follow from the general theory, so it is applicable only for log-density estimation procedure.

The results require some tools from linear algebra, technical lemmas on small bias and one result for deviation bounds for quadratic forms, which are presented in appendix.

2 Conditions, Level 1

We introduce four conditions: (\mathcal{C}) , (\mathcal{I}) , (\mathcal{L}_0) , (ED_0) , according to Spokoiny. These conditions are used to prove the theorems in the section 4. We are going to check these conditions in the section 5 after formulation of the main results.

2.1 Identification Condition

(\mathcal{I}) There exists a constant $\mathfrak{a} > 0$ such that

$$\mathfrak{a}^2 D_n^2 \succeq V_n^2 . \quad (2.1)$$

This condition will be checked with \mathfrak{a} close to 1. The exact value of \mathfrak{a} depends on h , if $h \rightarrow 0$ then $\mathfrak{a} \rightarrow 1$. In the essence, it depends only on the bias between polynomial basis and the true density function on the interval.

2.2 Local Identifiability Condition

(\mathcal{L}_0) There exists a constant $\delta_n(\mathbf{r}_0)$, depending on the sample size and concentration radius such that for each $\boldsymbol{\theta} \in \Theta_n(\mathbf{z})$ it holds:

$$\|\mathbf{I}_p - D_n^{-1} D_n^2(\boldsymbol{\theta}) D_n^{-1}\| \leq \delta_n(\mathbf{r}_0) . \quad (2.2)$$

This condition relates the matrices $D_n^2(\boldsymbol{\theta})$ and $D_n^2(\boldsymbol{\theta}^*)$ in terms of eigenvectors and eigenvalues. It is a standard tool for matrix comparison, and we shall see that many matrices that encounter in this article, obey the similar law.

2.3 Exponential Moment Condition

(ED_0) Let $\boldsymbol{\zeta} = V_n^{-1} \nabla L - \mathbf{E} V_n^{-1} \nabla L$. There exist constants $\mathfrak{g} > 0$ and $\nu_0 > 0$ such that for $\forall \boldsymbol{\gamma} \in \mathbb{R}^p$:

$$\log \mathbf{E} \exp \left(\lambda \frac{\boldsymbol{\gamma}^\top \boldsymbol{\zeta}}{\|\boldsymbol{\gamma}\|} \right) \leq \frac{\nu_0^2 \lambda^2}{2} \quad \forall \lambda: |\lambda| \leq \mathfrak{g} \quad (2.3)$$

Both ν_0 and \mathfrak{g} enter final quantile function and probability, so it is possible to perform some nontrivial optimization to obtain some sharper bounds. This condition can be satisfied with finite ν_0 and $\mathfrak{g} = \infty$ (we don't give proof of this fact, though it can be deduced from how we check this condition in section 5), but it is better to choose $\nu_0 \approx \sqrt{p}$ and some finite \mathfrak{g} , depending on the sample size n and bandwidth h .

2.4 Concentration Condition

(\mathcal{C}) The concentration radius \mathbf{r}_0 satisfies the inequality

$$\mathbf{r}_0(1 - \delta_n(\mathbf{r}_0)) \geq 2\zeta(p, \mathbf{z}) . \quad (2.4)$$

This condition is mainly an implicit rule for defining \mathbf{r}_0 . It is implicit because the constant δ depends on \mathbf{r}_0 , and this inequality can be satisfied for large enough n , because $\delta_n = O((nh^d)^{-1/2})$.

We shall see that it is possible to choose particular \mathbf{r}_0 if the effective sample size is not very small. Otherwise, we should correct the quantile function $\zeta(p, \mathbf{z})$ which will lead to different probability in concentration theorem.

3 Constants, Level 2

In order to satisfy these conditions, we need to establish the relationships between the objects from section 1.1. We are going to reformulate the conditions from the section 2 in terms of the basis Ψ and true density function $f(x)$.

3.1 Small Oscillation Condition

Let $f(x)$ be a true density function. There exists a constant $c_{f,h}$ such that:

$$\left| 1 - \frac{f(x)}{f(x_0)} \right| \leq c_{f,h} \ , \quad \forall |x - x_0| \leq h \ . \quad (3.1)$$

It may seem that condition is rather crude, because in the case of polynomial basis we are estimating not only the value of the function $f(x_0)$, but also its derivatives, that are contained in the vector θ^\bullet . The correct estimation procedure should lead to correct derivatives. But the influence of this constant $c_{f,h}$, as we will see later, is not very large. The bias actually is more important, which is of order $O(h^p)$ for polynomial basis in one-dimensional case.

3.2 Small Bias Condition

There exists a vector θ^\bullet and a constant $B_{p,h}$ such that $\forall t \in [-1, 1]$ it holds:

$$B_{p,h} \geq \exp \left(\varphi(x_0 + th) - \Psi^\top(t) \theta^\bullet \right) \ , \quad (3.2)$$

where $\varphi(x) = \log f(x)$, with $f(x)$ as a true density function.

In case of one-dimensional polynomial basis $\Psi(t)$ if the function $\varphi(x) = \log f(x)$ is smooth enough, the constant $B_{p,h}$ is of order $1 + O(h^p)$ and can be bounded by

$$\log B_{p,h} \leq \frac{h^p}{p!} \max_{x \in U_h(x_0)} \varphi^{(p)}(x) \ . \quad (3.3)$$

In d -dimensional case, in order to make bias of order $O(h^p)$, we need to take $\binom{p+d}{d} - 1$ elements of the basis, for example in local quadratic fitting for two-dimensional space, $\Psi(\mathbf{t}) = (1, t_1, t_2, t_1^2, t_1 t_2, t_2^2)$.

3.3 Curve Optimization Condition

This condition is completely defined by the model and can be calculated by the statistician. We require that there exists finite constant \mathbf{c}_1 such that

$$\mathbf{c}_1^2 = \sup_{t \in [-1, 1]} \Psi^\top(t) \left[\int_{-1}^1 K(\tau) \Psi(\tau) \Psi^\top(\tau) d\tau \right]^{-1} \Psi(t) \ . \quad (3.4)$$

The constant \mathbf{c}_1 depends on basis, and is computable. In case of one-dimensional polynomial basis and indicator kernel it equals to $\mathbf{c}_1^2 = p^2/2$. The reader can

check, for example, that in two-dimensional ($d = 2$) quadratic case $\Psi(t) = (1, t_1, t_2, t_1^2, t_1 t_2, t_2^2)$ with indicator kernel this constant is well-defined and equals to $13/2$.

We can introduce another constant

$$\mathfrak{c}_2^2 = \sup_{t \in [-1, 1]} K(t)^2 \Psi^\top(t) \left[\int_{-1}^1 K(\tau) \Psi(\tau) \Psi^\top(\tau) d\tau \right]^{-1} \Psi(t) , \quad (3.5)$$

where it clearly holds $\mathfrak{c}_2 \leq \mathfrak{c}_1$. In order to choose the “best model”, both constants should be bounded from above as better as possible.

3.4 Small Bandwidth Condition

Here we define ϕ_1 and ϕ_2 , which depend on the true density value $f_0 = f(x_0)$, and also on oscillation and bias constants $c_{f,h}$ and $B_{p,h}$, but in a given explicit way:

$$\begin{aligned} \phi_1^2 &= 2 \int_{-1}^1 K(\tau) d\tau \cdot \left((1 \pm 1) c_{f,h} \log f_0 \mp c_{f,h} + (1 \pm c_{f,h}) \log(1 \pm c_{f,h}) \right) , \\ \phi_2^2 &= \int_{-1}^1 K(\tau) d\tau \cdot f_0^3 (c_{f,h} - \log B_{p,h})^2 , \end{aligned} \quad (3.6)$$

where the “ \pm ” sign stands for maximum of the two expressions with “ $-$ ” and “ $+$ ” respectively. We require that $\mathfrak{c}_1 \phi_1 < \frac{\sqrt{5}-1}{2} \approx 0.618$ and $\mathfrak{c}_1 \phi_2 < 1$, this condition arises in the proof of theorem 4, and also in check of the conditions (\mathcal{L}_0) , (\mathcal{C}) , lemmas 2, 3.

When $h \rightarrow 0$, this condition is fulfilled automatically, but this condition can also serve as an approximate strategy for choosing \hat{h} if we know the estimates for $\hat{f}(x_0)$ and $\hat{f}'(x_0)$.

3.5 Effective Sample Size Condition

The lower bound on effective sample size is given as lemma 3 and requires that

$$\sqrt{nh^d} \geq f(x_0) \frac{4\mathfrak{c}_1 \zeta(p, \mathbf{z})}{\log 3/2 \sqrt{1 - \mathfrak{c}_1 \phi_1}} . \quad (3.7)$$

However, in low-density regions where $f(x_0) \approx 0$ this approach becomes inconsistent. Discussion on this issue is also provided after lemma 3. When the effective sample size is too small, the results can be modified to remain valid, but with lower probabilities and quantile values.

4 Main Theorems

4.1 Concentration Result

Theorem 1. *Let the conditions (\mathcal{I}) , (\mathcal{L}_0) , (\mathcal{C}) , (ED_0) be satisfied with some constants \mathfrak{a} , ν_0 , \mathfrak{g} , $\mathbf{r}_0(\mathbf{z})$. Let*

$$\Theta_n(\mathbf{z}) = \left\{ \boldsymbol{\theta} : \|D_n(\boldsymbol{\theta} - \boldsymbol{\theta}^*)\| \leq \mathbf{r}_0(\mathbf{z}) \right\} , \quad (4.1)$$

$$\mathbf{r}_0(\mathbf{z}) = 4\mathfrak{a} \cdot \nu_0(\sqrt{p} + \sqrt{2\mathbf{z}}), \quad \mathbf{z} \leq \mathfrak{g}^2/4 . \quad (4.2)$$

Then

$$\mathbf{P}\left(\tilde{\boldsymbol{\theta}} \notin \Theta_n(\mathbf{z})\right) \leq 2e^{-\mathbf{z}} + 8.4e^{-\mathfrak{g}^2/4} . \quad (4.3)$$

Remark 1. There is a condition in the theorem that $\mathbf{z} \leq \mathfrak{g}^2/4$. In fact, it is not very restrictive because \mathfrak{g}^2 is of order nh^d . However, it is also possible to state the theorem for infinitely large values of \mathbf{z} , using a second version of the quantile function for sub-gaussian quadratic forms. The probability measure of the set $\Theta_n(\mathbf{z})$ will become $2e^{-\mathbf{z}}$.

Proof. Let $\tilde{\boldsymbol{\theta}} \notin \Theta_n(\mathbf{r}_0)$. Since $\tilde{\boldsymbol{\theta}}$ maximizes log-likelihood, we have

$$L(\tilde{\boldsymbol{\theta}}) \geq L(\boldsymbol{\theta}^*) . \quad (4.4)$$

Since $L(\boldsymbol{\theta})$ is concave in $\boldsymbol{\theta}$, there exists a point

$$\check{\boldsymbol{\theta}} = \lambda\tilde{\boldsymbol{\theta}} + (1 - \lambda)\boldsymbol{\theta}^*, \quad \lambda \in [0, 1] \quad (4.5)$$

with the properties

$$\|D_n(\check{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)\| = \mathbf{r}_0 , \quad L(\check{\boldsymbol{\theta}}) \geq L(\boldsymbol{\theta}^*) . \quad (4.6)$$

It is enough to show that with probability $1 - 2e^{-\mathbf{z}} - 8.4e^{-\mathfrak{g}^2/4}$ it holds

$$L(\boldsymbol{\theta}) - L(\boldsymbol{\theta}^*) < 0, \quad \forall \boldsymbol{\theta} \notin \Theta_n(\mathbf{z}) . \quad (4.7)$$

Let us represent log-likelihood in the form

$$L(\boldsymbol{\theta}) = S^\top \boldsymbol{\theta} - A(\boldsymbol{\theta}) . \quad (4.8)$$

Since $\nabla \mathbf{E}L(\boldsymbol{\theta}^*) = 0$,

$$\mathbf{E}S = \nabla A(\boldsymbol{\theta}^*) . \quad (4.9)$$

Therefore, for any $\boldsymbol{\theta}$ it holds:

$$\begin{aligned} L(\boldsymbol{\theta}) - L(\boldsymbol{\theta}^*) &= S^\top (\boldsymbol{\theta} - \boldsymbol{\theta}^*) - [A(\boldsymbol{\theta}) - A(\boldsymbol{\theta}^*)] \\ &= (S - \mathbf{E}S)^\top (\boldsymbol{\theta} - \boldsymbol{\theta}^*) - \end{aligned} \quad (4.10)$$

$$[A(\boldsymbol{\theta}) - A(\boldsymbol{\theta}^*) - (\boldsymbol{\theta} - \boldsymbol{\theta}^*)^\top \nabla A(\boldsymbol{\theta}^*)] . \quad (4.11)$$

Inspect the first summand:

$$(S - \mathbf{E}S)^\top (\boldsymbol{\theta} - \boldsymbol{\theta}^*) = [D_n^{-1}(S - \mathbf{E}S)]^\top D_n(\boldsymbol{\theta} - \boldsymbol{\theta}^*) \quad (4.12)$$

For vector $\boldsymbol{\xi} = D_n^{-1}\nabla L(\boldsymbol{\theta}^*) = D_n^{-1}(S - \mathbf{E}S)$ it follows by lemma 12 for \mathfrak{a} and ν_0 from conditions section 2 that:

$$\mathbf{P}(\|\boldsymbol{\xi}\| \geq \zeta(p, \mathbf{z})) \leq 2e^{-\mathbf{z}} + 8.4e^{-\mathfrak{g}^2/4}, \quad \zeta^2(p, \mathbf{z}) = \mathfrak{a}^2\nu_0^2(p + 2\sqrt{2p\mathbf{z}} + 2\mathbf{z}) . \quad (4.13)$$

Therefore,

$$\|S^\top(\boldsymbol{\theta} - \boldsymbol{\theta}^*)\| = \|\boldsymbol{\xi}^\top D_n(\boldsymbol{\theta} - \boldsymbol{\theta}^*)\| \leq \zeta(p, \mathbf{z}) \cdot \mathbf{r}_0 . \quad (4.14)$$

The second summand, by Taylor expansion, can be represened as

$$A(\boldsymbol{\theta}) - A(\boldsymbol{\theta}^*) - (\boldsymbol{\theta} - \boldsymbol{\theta}^*)^\top \nabla A(\boldsymbol{\theta}^*) = \frac{1}{2}(\boldsymbol{\theta} - \boldsymbol{\theta}^*)^\top \nabla^2 A(\bar{\boldsymbol{\theta}})(\boldsymbol{\theta} - \boldsymbol{\theta}^*) . \quad (4.15)$$

By condition (\mathcal{L}_0) with $\delta_n(\mathbf{r}_0)$ and $\bar{\boldsymbol{\theta}} \in \Theta_n(\mathbf{z})$ it follows that

$$\frac{1}{2}(\boldsymbol{\theta} - \boldsymbol{\theta}^*)^\top \nabla^2 A(\bar{\boldsymbol{\theta}})(\boldsymbol{\theta} - \boldsymbol{\theta}^*) \geq \frac{1 - \delta_n(\mathbf{r}_0)}{2} \|D_n(\boldsymbol{\theta} - \boldsymbol{\theta}^*)\|^2 . \quad (4.16)$$

Thus, with probatily at least $1 - 2e^{-\mathbf{z}} - 8.4e^{-\mathfrak{g}^2/4}$ it follows that

$$L(\check{\boldsymbol{\theta}}) - L(\boldsymbol{\theta}^*) \leq \zeta(p, \mathbf{z})\mathbf{r}_0 - \frac{1 - \delta_n(\mathbf{r}_0)}{2} \mathbf{r}_0^2 \leq 0 , \quad (4.17)$$

which is contradiction, according to the condition (\mathcal{C}) ,

$$\mathbf{r}_0(\mathbf{z})(1 - \delta_n(\mathbf{r}_0(\mathbf{z}))) \geq 2\zeta(p, \mathbf{z}) . \quad (4.18)$$

End of the proof of theorem 1.

4.2 Fisher Theorem

This theorem describes finite-sample approximation of the distribution of the estimate $\tilde{\boldsymbol{\theta}}$ in terms of D_n and score vector $\boldsymbol{\xi}$.

Theorem 2. *Let the conditions (\mathcal{C}) , (\mathcal{I}) , (\mathbf{ED}_0) , (\mathcal{L}_0) hold.*

Then for $\tilde{\boldsymbol{\theta}} \in \Theta_n(\mathbf{z})$ from theorem 1 with dominating probability

$$\mathbf{P}(\tilde{\boldsymbol{\theta}} \in \Theta_n(\mathbf{z})) \geq 1 - (2e^{-\mathbf{z}} + 8.4e^{-\mathfrak{g}^2/4}) \quad (4.19)$$

it holds

$$\|D_n(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^*) - \boldsymbol{\xi}\| \leq \diamond(n, \mathbf{z}) , \quad \mathbf{z} \leq \mathfrak{g}^2/4 , \quad (4.20)$$

where $\diamond(n, \mathbf{z}) = \mathbf{r}_0(\mathbf{z}) \cdot \delta_n(\mathbf{r}_0)$, and \mathbf{r}_0 is defined by (4.2).

Remark 2. The vector $(nh)^{-1/2}\boldsymbol{\xi}_n$ is asymptotically standard normal. Following the classical statistics, the difference between the centered parameter and the standard normal random variable is of order $(nh^d)^{-1/2}$:

$$\|d_0(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^*) - \mathcal{N}(0, 1)\| = O((nh^d)^{-1/2}) . \quad (4.21)$$

The Fisher theorem is the asymptotic refinement of the Central Limit Theorem. Indeed, asymptotic behavior of the term $\delta_n(\mathbf{r}_0)$ is the following: with $nh \rightarrow \infty$, we have $\delta_n(\mathbf{r}_0) \rightarrow 0$, $\diamond(n, \mathbf{z}) = O((nh)^{-1/2})$.

While $\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^* = O((nh)^{-1/2})$, the Fisher result can be written in the form:

$$\left\| d_0(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^*) - (nh^d)^{-1/2}\boldsymbol{\xi}_n \right\| = O((nh^d)^{-1}) . \quad (4.22)$$

Proof. The principle step is a bound on the local linear approximation of the stochastic part of the gradient $\nabla L(\boldsymbol{\theta})$. Although $\boldsymbol{\xi}$ is random, it can be shown that $\boldsymbol{\xi}$ depends only on $\tilde{\boldsymbol{\theta}}$.

Indeed, since $\nabla L(\tilde{\boldsymbol{\theta}})$ is zero, and stochastic part of L is linear on $\boldsymbol{\theta}$,

$$\boldsymbol{\xi}(\boldsymbol{\theta}^*) = D_n^{-1} \nabla L(\boldsymbol{\theta}^*) = D_n^{-1} [\nabla L(\boldsymbol{\theta}^*) - \nabla L(\tilde{\boldsymbol{\theta}})] = D_n^{-1} [\nabla \text{EL}(\boldsymbol{\theta}^*) - \nabla \text{EL}(\boldsymbol{\theta})|_{\boldsymbol{\theta}=\tilde{\boldsymbol{\theta}}}] . \quad (4.23)$$

Next, we can bound the norm of $(D_n(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^*) - \boldsymbol{\xi})$ by multiplying it by an arbitrary vector \mathbf{u} of unit norm:

$$\begin{aligned} \mathbf{u}^\top [D_n(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^*) - \boldsymbol{\xi}] &= \mathbf{u}^\top \left[D_n(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^*) - D_n^{-1}(\nabla \text{EL}(\boldsymbol{\theta}^*) - \nabla \text{EL}(\tilde{\boldsymbol{\theta}})) \right] \\ &= \mathbf{u}^\top \left[D_n(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^*) - D_n^{-1} D_n^2(\bar{\boldsymbol{\theta}}_{\mathbf{u}})(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^*) \right] \\ &= \mathbf{u}^\top \left[\mathbf{I}_p - D_n^{-1} D_n^2(\bar{\boldsymbol{\theta}}_{\mathbf{u}}) D_n^{-1} \right] (\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^*) \\ &\leq \|\mathbf{u}^\top (\mathbf{I}_p - D_n^{-1} D_n^2(\bar{\boldsymbol{\theta}}_{\mathbf{u}}) D_n^{-1})\| \cdot \|D_n(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)\| , \end{aligned}$$

where $\bar{\boldsymbol{\theta}}_{\mathbf{u}} = \lambda \tilde{\boldsymbol{\theta}} + (1 - \lambda) \boldsymbol{\theta}^*$, $\lambda \in [0, 1]$.

By theorem 1, with high probability it holds $\tilde{\boldsymbol{\theta}} \in \Theta_n(\mathbf{z})$. By condition (\mathcal{L}_0) , for each $\bar{\boldsymbol{\theta}} \in \Theta_n(\mathbf{z})$ it holds $\|\mathbf{I}_p - D_n^{-1} D_n^2(\bar{\boldsymbol{\theta}}) D_n^{-1}\| \leq \delta(\mathbf{r}_0)$, so we have

$$\|D_n(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^*) - \boldsymbol{\xi}\| \leq \mathbf{r}_0(\mathbf{z}) \delta(\mathbf{r}_0) . \quad (4.24)$$

End of the proof of theorem 2.

4.3 Wilks Theorem

Theorem 3 (Spokoiny, [1]). *Let the conditions (\mathcal{C}) , (\mathcal{I}) , (ED_0) , (\mathcal{L}_0) hold.*

Then for $\tilde{\boldsymbol{\theta}} \in \Theta_n(\mathbf{z})$ from theorem 1 with dominating probability

$$\mathbf{P} \left(\tilde{\boldsymbol{\theta}} \in \Theta_n(\mathbf{z}) \right) \geq 1 - (2e^{-\mathbf{z}} + 8.4e^{-\mathbf{z}^2/4}) \quad (4.25)$$

with $\diamond(n, \mathbf{z})$ from theorem 2 it holds:

$$\left| \sqrt{2L(\tilde{\boldsymbol{\theta}}, \boldsymbol{\theta}^*)} - \|D_n(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^\circ)\| \right| \leq 2\diamond(n, \mathbf{z}) \quad (4.26)$$

$$\left| \sqrt{2L(\tilde{\boldsymbol{\theta}}, \boldsymbol{\theta}^*)} - \|\boldsymbol{\xi}\| \right| \leq 3\diamond(n, \mathbf{z}) \quad (4.27)$$

Remark 3. The constant $\diamond(n, \mathbf{z})$ is familiar from the theorem 2, the theorem is an asymptotic refinement to behavior of likelihood when $nh^d \rightarrow \infty$.

The proof can be found in Spokoiny [1]. The theorem is valid under conditions, formulated in the abovementioned article, which are checked in the current text. \square

4.4 Accurate Small Bias Result

Theorem 4. Suppose that $\mathbf{c}_1\phi_1 \leq \frac{\sqrt{5}-1}{2}$ and $\mathbf{c}_2\phi_2 \leq 1$, $I_k = \int_{-1}^1 K(t)dt \leq 2^d$,

$$\varepsilon = \max\{\mathbf{c}_1\phi_1(1 - \mathbf{c}_1\phi_1)^{-1/2}, \mathbf{c}_1\phi_2\} . \quad (4.28)$$

Then it holds:

$$\|d_0(\boldsymbol{\theta}^\circ)(\boldsymbol{\theta}^* - \boldsymbol{\theta}^\bullet)\| \lesssim \sqrt{p}\sqrt{I_K}(1 - \varepsilon)^{-1}(1 + c_{f,h}) \cdot f(x_0) \cdot |B_{p,h} - 1| . \quad (4.29)$$

Remark 4. There are results of a kind $\boldsymbol{\theta}^* \approx \boldsymbol{\theta}^\circ$ and $\boldsymbol{\theta}^\circ \approx \boldsymbol{\theta}^\bullet$, in terms of curvature matrix, see lemmas 5 and 6. However, we cannot combine these results to obtain the final bound, because it is an asymptotic refinement of order $O(h^p)$ instead of $O(h)$. More precisely, the term $|B_{p,h} - 1|$ is of order $O(h^p)$, other terms are of order $1 + O(h)$, and the following approximate inequality holds:

$$\|d_0(\boldsymbol{\theta}^\circ)(\boldsymbol{\theta}^* - \boldsymbol{\theta}^\bullet)\| \lesssim \sqrt{2}f(x_0)|B_{p,h} - 1| . \quad (4.30)$$

Proof. From the conditions $\mathbf{c}_1\phi_1 < \frac{\sqrt{5}-1}{2}$ it follows that $\mathbf{c}_1\phi_1(1 - \mathbf{c}_1\phi_1)^{-1/2} < 1$. We will use this observation later. Combined with $\mathbf{c}_1\phi_2 < 1$ this allows to claim that $\varepsilon < 1$.

Let $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2 \in [\boldsymbol{\theta}^*, \boldsymbol{\theta}^\bullet]$. If there is a point $\boldsymbol{\theta}^t = t\boldsymbol{\theta}^\bullet + (1 - t)\boldsymbol{\theta}^*$, $t \in [0, 1]$, then there is a representation

$$\boldsymbol{\theta}^t - \boldsymbol{\theta}^\circ = t\boldsymbol{\theta}^\bullet + (1 - t)\boldsymbol{\theta}^* - t\boldsymbol{\theta}^\circ - (1 - t)\boldsymbol{\theta}^\circ = t(\boldsymbol{\theta}^\bullet - \boldsymbol{\theta}^\circ) + (1 - t)(\boldsymbol{\theta}^* - \boldsymbol{\theta}^\circ) . \quad (4.31)$$

Hence, we have a triangle inequality in terms of curvature matrix $d_0(\boldsymbol{\theta}^\circ)$:

$$\begin{aligned} \|d_0(\boldsymbol{\theta}^\circ)(\boldsymbol{\theta}^t - \boldsymbol{\theta}^\circ)\| &\leq t\|d_0(\boldsymbol{\theta}^\circ)(\boldsymbol{\theta}^\bullet - \boldsymbol{\theta}^\circ)\| + (1 - t)\|d_0(\boldsymbol{\theta}^\circ)(\boldsymbol{\theta}^* - \boldsymbol{\theta}^\circ)\| \\ &\leq \max\{\|d_0(\boldsymbol{\theta}^\circ)(\boldsymbol{\theta}^* - \boldsymbol{\theta}^\circ)\|, \|d_0(\boldsymbol{\theta}^\circ)(\boldsymbol{\theta}^\bullet - \boldsymbol{\theta}^\circ)\|\} . \end{aligned} \quad (4.32)$$

Therefore, according to lemmas 5 and 6, at each of the points $\boldsymbol{\theta}^t \in \{\boldsymbol{\theta}_1, \boldsymbol{\theta}_2\}$ it holds

$$\begin{aligned} \|\mathbf{I}_p - D(\boldsymbol{\theta}^t)D^{-2}(\boldsymbol{\theta}^\circ)D(\boldsymbol{\theta}^t)\| &\leq \exp\left(\boldsymbol{\Psi}^\top(t)d_0^{-1}d_0(\boldsymbol{\theta}^t - \boldsymbol{\theta}^\circ)\right) - 1 \\ &\lesssim \|\boldsymbol{\Psi}^\top(t)d_0^{-1}\| \cdot \|d_0(\boldsymbol{\theta}^t - \boldsymbol{\theta}^\circ)\| \\ &\leq \mathbf{c}_1 \cdot \max\{\|d_0(\boldsymbol{\theta}^\circ)(\boldsymbol{\theta}^* - \boldsymbol{\theta}^\circ)\|, \|d_0(\boldsymbol{\theta}^\circ)(\boldsymbol{\theta}^\bullet - \boldsymbol{\theta}^\circ)\|\} \\ &\leq \mathbf{c}_1 \cdot \max\left\{\phi_1(1 - \mathbf{c}_1\phi_1)^{-1/2}, \phi_2\right\} . \end{aligned}$$

Since $\varepsilon = \mathbf{c}_1 \max\{\phi_1(1 - \mathbf{c}_1\phi_1)^{-1/2}, \phi_2\}$, we obtain:

$$\|\mathbf{I}_p - D(\boldsymbol{\theta}^t)D^{-2}(\boldsymbol{\theta}^\circ)D(\boldsymbol{\theta}^t)\| \leq \varepsilon . \quad (4.33)$$

Next, we construct a bound using two Taylor expansions. Let $g(\boldsymbol{\theta}) = (nh)^{-1}\mathbf{E}L(\boldsymbol{\theta})$.

$$g(\boldsymbol{\theta}^*) = g(\boldsymbol{\theta}^\bullet) + \nabla g(\boldsymbol{\theta}^\bullet)^\top(\boldsymbol{\theta}^* - \boldsymbol{\theta}^\bullet) + \frac{1}{2}(\boldsymbol{\theta}^* - \boldsymbol{\theta}^\bullet)^\top d_0^2(\boldsymbol{\theta}_1)(\boldsymbol{\theta}^* - \boldsymbol{\theta}^\bullet) \quad (4.34)$$

$$g(\boldsymbol{\theta}^\bullet) = g(\boldsymbol{\theta}^*) + \nabla g(\boldsymbol{\theta}^*)^\top(\boldsymbol{\theta}^\bullet - \boldsymbol{\theta}^*) + \frac{1}{2}(\boldsymbol{\theta}^\bullet - \boldsymbol{\theta}^*)^\top d_0^2(\boldsymbol{\theta}_2)(\boldsymbol{\theta}^\bullet - \boldsymbol{\theta}^*) \quad (4.35)$$

Adding the two expressions, we obtain

$$\begin{aligned} (1 - \varepsilon)\|d_0(\boldsymbol{\theta}^\circ)(\boldsymbol{\theta}^* - \boldsymbol{\theta}^\bullet)\|^2 &\leq (\boldsymbol{\theta}^* - \boldsymbol{\theta}^\bullet)^\top \frac{d_0^2(\boldsymbol{\theta}_1) + d_0^2(\boldsymbol{\theta}_2)}{2}(\boldsymbol{\theta}^* - \boldsymbol{\theta}^\bullet) \\ &\leq \|\nabla(g(\boldsymbol{\theta}^\bullet) - g(\boldsymbol{\theta}^*))^\top(\boldsymbol{\theta}^* - \boldsymbol{\theta}^\bullet)\| \end{aligned} \quad (4.36)$$

Then latter, by Cauchy inequality, can be bounded by

$$\begin{aligned} \|\nabla(g(\boldsymbol{\theta}^\bullet) - g(\boldsymbol{\theta}^*))^\top(\boldsymbol{\theta}^* - \boldsymbol{\theta}^\bullet)\| &\leq \|d_0(\boldsymbol{\theta}^\circ)(\boldsymbol{\theta}^* - \boldsymbol{\theta}^\circ)\| \times \\ &\quad \|d_0(\boldsymbol{\theta}^\circ)^{-1}(\nabla g(\boldsymbol{\theta}^*) - \nabla g(\boldsymbol{\theta}^\bullet))\| . \end{aligned} \quad (4.37)$$

Therefore, after cancelling $\|d_0(\boldsymbol{\theta}^\circ)(\boldsymbol{\theta}^* - \boldsymbol{\theta}^\bullet)\|$ from both sides, we have, according to the lemma 7:

$$\begin{aligned} \|d_0(\boldsymbol{\theta}^\circ)(\boldsymbol{\theta}^* - \boldsymbol{\theta}^\bullet)\| &\leq (1 - \varepsilon)^{-1}\|d_0^{-1}(\boldsymbol{\theta}^\circ)(\nabla g(\boldsymbol{\theta}^*) - \nabla g(\boldsymbol{\theta}^\bullet))\| \\ &\leq (1 - \varepsilon)^{-1}|1 - B_{p,h}| \cdot \sqrt{ph^{-d}\text{pr}_2(x_0)} , \end{aligned} \quad (4.38)$$

where

$$\sqrt{h^{-d}\text{pr}_2(x_0)} \leq \sqrt{\int_{-1}^1 K(t)f(x_0)^2(1 + c_{f,h})^2 dt} \leq \sqrt{I_K}f(x_0)(1 + c_{f,h}) . \quad (4.39)$$

End of the proof of theorem 4.

5 Checking Conditions

5.1 Identification Condition (I)

Lemma 1. *Let the conditions from theorem 4 hold. Then the constant \mathfrak{a} in condition (I):*

$$\mathfrak{a}^2 D_n^2 \succeq V_n^2 \quad (5.1)$$

can be bounded above by

$$\mathfrak{a}^2 \leq \sup_{|x-x_0| \leq h} \frac{K((x-x_0)/h)f(x)}{\exp(\Psi^\top \theta^*)} \quad (5.2)$$

$$\leq B_{p,h} \exp(\mathfrak{c}_1 \cdot I_K^{1/2} (1-\varepsilon)^{-1} (1+c_{f,h}) f(x_0) |B_{p,h} - 1|) , \quad (5.3)$$

where $I_K = \int_{-1}^1 K(t)dt$, $\varepsilon = \max\{\mathfrak{c}_1 \phi_1 (1 - \mathfrak{c}_1 \phi_1)^{-1/2}, \mathfrak{c}_1 \phi_2\}$.

Remark 5. In case of one-dimensional local polynomial estimation with indicator kernel, the quantity $|B_{p,h} - 1|$ is of order $O(h^p)$, the multiples $(1-\varepsilon)$, $(1+c_{f,h})$ are of order 1. Therefore, for h small enough, $h < 1$, we have

$$\mathfrak{a}^2 \lesssim B_{p,h} (1 + \sqrt{p I_K} \mathfrak{c}_1 f(x_0) \cdot O(h^p)) \approx 1 + \sqrt{p I_K} \mathfrak{c}_1 f(x_0) \cdot |B_{p,h} - 1| . \quad (5.4)$$

Proof. Firstly bound $V_n^2 = \text{Var } \nabla L$.

$$\begin{aligned} \frac{1}{n} (V_n^2)_{ij} &= \int_{\mathcal{X}} K^2 \Psi_i \Psi_j f(x) dx - \int_{\mathcal{X}} K \Psi_i f(x) dx \int_{\mathcal{X}} K \Psi_j f(x) dx \\ &\preceq \int_{\mathcal{X}} K^2 \Psi_i \Psi_j f(x) dx. \end{aligned}$$

We use the fact that the second summand is minus non-negative definite matrix with rank 1. Since both matrices D_n^2 and first summand of V_n^2 can be represented in the form suitable for lemma 8, we can apply it and bound the maximal eigenvalue of $D_n^{-1} V_n^2 D_n^{-1}$.

Recall that

$$D_n^2 = \int_{\mathcal{X}} K \Psi \Psi^\top \exp(\Psi^\top(x) \theta^*) dx . \quad (5.5)$$

In terms of lemma 8 their diagonal operators are, correspondingly, $K(t) \exp(\Psi^\top(t) \theta^*)$ and $K^2(t) f(x)$. So, \mathfrak{a}^2 can be estimated with

$$\mathfrak{a}^2 \leq \sup_{x \in U_h(x_0)} \frac{K(t) f(x)}{\exp(\Psi^\top(t) \theta^*)} . \quad (5.6)$$

Then we use trivial bound $K(t) \leq 1$. It is possible to write

$$\sup_{|x_0-x| \leq h} \frac{f(x)}{\exp(\Psi^\top \theta^*)} \leq \exp(\Psi^\top(t) (\theta^\bullet - \theta^*) + \log B_{p,h}) . \quad (5.7)$$

Using the result of theorem 4 we obtain

$$\mathfrak{a}^2 \leq B_{p,h} \exp(\Psi^\top(t) d_0^{-1}(\theta^\circ) d_0(\theta^\circ) (\theta^\bullet - \theta^*)) \quad (5.8)$$

$$\leq B_{p,h} \exp(\mathfrak{c}_1 \cdot \sqrt{p} \sqrt{I_K} (1-\varepsilon)^{-1} (1+c_{f,h}) f(x_0) |B_{p,h} - 1|) . \quad (5.9)$$

End of the proof of lemma 1.

5.2 Local Identifiability Condition (\mathcal{L}_0)

Lemma 2. For all $\boldsymbol{\theta} \in \Theta_n(\mathbf{z})$ local identifiability condition

$$\|\mathbf{I}_p - D^{-1}D^2(\boldsymbol{\theta})D^{-1}\| \leq \delta_n(\mathbf{r}_0(\mathbf{z})) \quad (5.10)$$

holds with the constant

$$\delta_n(\mathbf{r}_0) \leq \exp\left(\frac{\mathbf{c}_1 \mathbf{r}_0}{\sqrt{1 - \mathbf{c}_1 \phi_1} \sqrt{f(x_0)nh^d}}\right) - 1 . \quad (5.11)$$

Remark 6. When effective sample size is large, and h is small, the above expression is equivalent to $\delta_n \lesssim \frac{\mathbf{c}_1 \mathbf{r}_0}{\sqrt{f(x_0)nh^d}}$. We will see later that r_0 can be chosen as $4\zeta(p, \mathbf{z})$.

Proof. Maximal absolute eigenvalue of $(\mathbf{I} - X)$ is equal to $\max(|\lambda_{\min}(X) - 1|, |\lambda_{\max}(X) - 1|)$. From lemma 8 it follows that $\lambda(D_n^{-1}D_n^2(\boldsymbol{\theta})D_n^{-1})$ belongs to the interval

$$\left[\min_{x \in U_h(x_0)} \exp(\boldsymbol{\Psi}(x)^\top(\boldsymbol{\theta} - \boldsymbol{\theta}^*)), \max_{x \in U_h(x_0)} \exp(\boldsymbol{\Psi}(x)^\top(\boldsymbol{\theta} - \boldsymbol{\theta}^*)) \right] . \quad (5.12)$$

Let $\mathbf{v} = \boldsymbol{\theta} - \boldsymbol{\theta}^*$. Note that $\boldsymbol{\Psi}^\top \mathbf{v} = \boldsymbol{\Psi}^\top D_n^{-1}(\boldsymbol{\theta}^*)D_n(\boldsymbol{\theta}^*)\mathbf{v}$ and consequently, the matrices $D_n^2(\boldsymbol{\theta}^*)$ and $D_n^2(\boldsymbol{\theta}^\circ)$ are related through lemma 5. Therefore,

$$\|\boldsymbol{\Psi}^\top D_n^{-1}(\boldsymbol{\theta}^*)\|^2 \leq (f(x_0)nh^d)^{-1} \mathbf{c}_1^2 \cdot (1 - \mathbf{c}_1 \phi_1)^{-1} , \quad \|D_n(\boldsymbol{\theta}^*)\mathbf{v}\| \leq \mathbf{r}_0 , \quad (5.13)$$

and by Cauchy inequality the constant δ_n is bounded by

$$\delta_n(\mathbf{r}_0) \leq \exp\left(\frac{\mathbf{c}_1 \mathbf{r}_0}{\sqrt{1 - \mathbf{c}_1 \phi_1} \sqrt{f(x_0)nh^d}}\right) - 1 . \quad (5.14)$$

End of the proof of lemma 2.

5.3 Concentration Condition (\mathcal{C})

Lemma 3. Under condition

$$\sqrt{nh^d} \geq f(x_0) \frac{4\mathbf{c}_1 \zeta(p, \mathbf{z})}{\log 3/2 \sqrt{1 - \mathbf{c}_1 \phi_1}} , \quad (5.15)$$

the concentration condition (\mathcal{C}) holds:

$$\mathbf{r}_0(\mathbf{z})(1 - \delta_n(\mathbf{r}_0)) \geq 2\zeta(p, \mathbf{z}) . \quad (5.16)$$

Proof. Note that $\delta_n(\mathbf{r}_0) \rightarrow 0$ when $nh^d \rightarrow 0$. We will need $nh^d > N_0$ such that $\delta_n(\mathbf{r}_0) \leq 1/2$. This will allow us to take $\mathbf{r}_0(\mathbf{z}) = 4\zeta(p, \mathbf{z})$. The condition turns into

$$\exp\left(\frac{\mathbf{c}_1 \mathbf{r}_0}{\sqrt{1 - \mathbf{c}_1 \phi_1} \sqrt{f(x_0) nh^d}}\right) \leq 3/2, \quad (5.17)$$

which turns into

$$\sqrt{nh^d} \geq (1 - \mathbf{c}_1 \phi_1)^{-1/2} f(x_0) \frac{4\mathbf{c}_1 \zeta(p, \mathbf{z})}{\log 3/2} \quad (5.18)$$

End of the proof of lemma 3.

Remark 7. When the density is small, there will be no concentration and the sample size will be too small for this condition. We can redefine $\zeta_n(p, \mathbf{z})$ as maximal value that satisfies the concentration condition

$$\zeta_n(p, \mathbf{z}) = \sqrt{1 - \mathbf{c}_1 \phi_1} \frac{\sqrt{nh^d} \log 3/2}{4\mathbf{c}_1 f(x_0)}, \quad \mathbf{r}_n(\mathbf{z}) = 4\zeta_n(p, \mathbf{z}), \quad (5.19)$$

and the probability in theorems 1 and 2 becomes

$$\mathbf{P}(\tilde{\boldsymbol{\theta}} \notin \Theta_n(\mathbf{z})) = \mathbf{P}(\|\boldsymbol{\xi}\| > \zeta_n(p, \mathbf{z})) . \quad (5.20)$$

Finally, for small $\sqrt{nh^d}$ the concentration theorem loses its “concentration” and can be stated in the following form:

$$\|d_0(\boldsymbol{\theta}^*)(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)\| \leq \sqrt{1 - \mathbf{c}_1 \phi_1} \frac{\log 3/2}{4\mathbf{c}_1 f(x_0)}, \quad (5.21)$$

where right-hand side is no longer of order $(nh^d)^{-1/2}$.

5.4 Exponential Moment Condition (ED_0)

Lemma 4. Let $\boldsymbol{\zeta} = \nabla L - \mathbf{E} \nabla L$. Consider the function

$$M(\lambda, \gamma) = \log \mathbf{E} \exp(\lambda \boldsymbol{\gamma}^\top V_n^{-1} \boldsymbol{\zeta}) \quad (5.22)$$

For all $|\lambda| \leq \mathfrak{g}$ the following inequality holds:

$$M(\lambda, \gamma) \leq \frac{\nu_0^2 \lambda^2}{2}, \quad \nu_0^2 = p + \frac{16\mathfrak{g} C_{V,f}^3}{\sqrt{nh^{3d}}}, \quad (5.23)$$

where $C_{V,f}$ is defined in lemma 11, and satisfies

$$C_{V,f}^2 \leq (1 - c_{f,h})^{-1} f_0^{-1} \mathbf{c}_2^2 + \frac{h^d}{1 - \text{pr}_1(x_0)}, \quad \text{pr}_1(x_0) = \int_{-1}^1 K((x - x_0)/h) f(x) dx . \quad (5.24)$$

Proof. Since ζ can be represented as a sum of i.i.d. random variables

$$\zeta = \sum_{i=1}^n \zeta_i = \sum_{i=1}^n (K_i \Psi_i - \mathbf{E} K_i \Psi_i) , \quad (5.25)$$

the function $M(\lambda, \gamma)$ can be also rewritten as

$$M(\lambda, \gamma) = n \log \mathbf{E} \exp(\lambda \gamma^\top V_n^{-1} \zeta_1) . \quad (5.26)$$

Consider Taylor expansion of degree 3 at $\lambda = 0$ for $M(\lambda, \gamma)$: there exists $\bar{\lambda} \in [0, \lambda]$ such that

$$M(\lambda, \gamma) = M(0) + \lambda M'(0, \gamma) + \frac{\lambda^2}{2} M''(0, \gamma) + \frac{\lambda^3}{6} M'''(\bar{\lambda}, \gamma) \quad (5.27)$$

Denote $u = \gamma^\top V_n^{-1} \zeta$, $u_1 = \gamma^\top V_n^{-1} \zeta_1$ for brevity. Careful differentiation gives us that

$$M'(\lambda) = \frac{\mathbf{E}(u \exp(\lambda u))}{\mathbf{E} \exp(\lambda u)}, \quad M''(\lambda) = \frac{\mathbf{E}(u^2 \exp(\lambda u)) \mathbf{E} \exp(\lambda u) - (\mathbf{E} u \exp(\lambda u))^2}{(\mathbf{E} \exp(\lambda u))^2} , \quad (5.28)$$

$$\frac{1}{n} M'''(\lambda) = \frac{\mathbf{E}(u_1^3 \exp(\lambda u_1))}{\mathbf{E} \exp(\lambda u_1)} - \frac{3 \mathbf{E}(u_1^2 \exp(\lambda u_1)) \mathbf{E}(u_1 \exp(\lambda u_1))}{(\mathbf{E} \exp \lambda u_1)^2} + \frac{2(\mathbf{E} u_1 \exp(\lambda u_1))^3}{(\mathbf{E} \exp(\lambda u_1))^3} , \quad (5.29)$$

and after substituting $\lambda = 0$ we obtain $M(0) = M'(0) = 0$, $M''(0) = \mathbf{E} u^2$. Then

$$M''(0) \leq \mathbf{E} \sup_{\|\gamma\|=1} (\gamma^\top V_n^{-1} \zeta)^2 = \mathbf{E} \zeta^\top V_n^{-2} \zeta = \text{Tr } \mathbf{E} \zeta \zeta^\top V_n^{-2} = p . \quad (5.30)$$

Let us show that $|u_1|^2 \leq \frac{4C_{V,f}^2}{nh^d}$, where $C_{V,f}$ is defined in lemma 11. First, we will show that the square of uncentered random variable $\gamma^\top V_n^{-1} K(T_1) \Psi(T_1)$ is bounded:

$$|\gamma^\top V_n^{-1} K(T_1) \Psi(T_1)|^2 \leq K(t)^2 \Psi(t)^\top V_n^{-2} \Psi(t) \leq \frac{1}{nh^d} C_{V,f}^2 . \quad (5.31)$$

Then, the centered variable is naturally bounded by twice the bound of noncentered, therefore the bound for square multiplies 4 times.

This observation allows to obtain the bound for $M'''(\lambda)$, using the fact that if random variable X is bounded by $|X| \leq c$, then $\mathbf{E} X Y \leq c \cdot \mathbf{E} |Y|$:

$$\frac{1}{n} M'''(\lambda) \leq 6 \cdot \left(\frac{2C_{V,f}}{\sqrt{nh^d}} \right)^3 . \quad (5.32)$$

Thus, we obtained the bound for the whole expression $M(\lambda, \gamma)$:

$$M(\lambda, \gamma) \leq \frac{\lambda^2}{2} p + \lambda^3 \cdot \frac{8C_{V,f}^3 n}{(\sqrt{nh^d})^3} = \frac{\lambda^2}{2} \left(p + 2\lambda \frac{8C_{V,f}^3}{\sqrt{nh^{3d}}} \right) \leq \frac{\lambda^2}{2} \left(p + \frac{16gC_{V,f}^3}{\sqrt{nh^{3d}}} \right) . \quad (5.33)$$

End of the proof of lemma 4.

A Technical Results

A.1 Small Bias Results for $\theta^\circ, \theta^\bullet, \theta^*$.

Lemma 5. Let $\mathbf{c}_1\phi_1 < 1$, $f_0 = f(x_0)$, where \mathbf{c}_1 is given by (3.4),

$$\phi_1^2 = 2I_k f_0 I_K \left((1 \pm 1)c_{f,h} \log f_0 \mp c_{f,h} + (1 \pm c_{f,h}) \log(1 \pm c_{f,h}) \right) , \quad (\text{A.1})$$

$$I_k = \int_{-1}^1 K(t) dt \leq 2^d , \quad (\text{A.2})$$

where the “ \pm ” sign stands for the maximum of two expressions with plus and minus. Then the following holds:

$$\|\mathbf{I}_p - D_n(\theta^*) D_n^{-2}(\theta^\circ) D_n(\theta^*)\| \lesssim \mathbf{c}_1 \phi_1 (1 - \mathbf{c}_1 \phi_1)^{-1} , \quad (\text{A.3})$$

$$\|d_0(\theta^\circ)(\theta^\circ - \theta^*)\|^2 \lesssim f_0 \phi_1^2 (1 - \mathbf{c}_1 \phi_1)^{-1} . \quad (\text{A.4})$$

Remark 8. The quantity ϕ_1^2 is proportional to $|2c_{f,h} \log f_0|$ which is of order $O(h)$, so with $h \rightarrow 0$ it holds $\|\mathbf{I}_p - D(\theta^*) D^{-2}(\theta^\circ) D(\theta^*)\| = O(h^{1/2})$, $\theta^\circ - \theta^* = O(h)$.

Proof. Consider the expectation of log-likelihood under the true measure:

$$\mathbf{E}L(\theta) = nh^d \left(\int_{-1}^1 K(t) \Psi(t)^\top \theta \cdot f(x_0 + th) dt - \int_{-1}^1 K(t) \exp(\Psi(t)^\top \theta) dt \right)$$

We would like to prove that $\theta^* \approx \theta^\circ$.

Denote $g(\theta) = (nh^d)^{-1} \mathbf{E}L(\theta)$, $f_0 = f(x_0)$.

Then

$$\begin{aligned} g(\theta^\circ) &= \int_{-1}^1 K(t) f(x_0 + ht) \cdot \log f_0 dt - \int_{-1}^1 K(t) f_0 dt \\ &\geq f_0 I_k \left((1 - c_{f,h}) \log f_0 - 1 \right) . \end{aligned}$$

From the other side, since for each $c > 0$ holds

$$cx - \exp(x) \leq c \log c - c , \quad (\text{A.5})$$

then it holds

$$\begin{aligned} g(\theta^*) &= \int_{-1}^1 K(t) \left(\Psi^\top \theta f(x_0 + ht) - \exp(\Psi^\top \theta) \right) dt \\ &\leq \int_{-1}^1 K(t) f(x_0 + ht) (\log f(x_0 + ht) - 1) dt . \end{aligned} \quad (\text{A.6})$$

Since the function $\varphi(\tau) = \tau(\log \tau - 1)$ is unimodal, for any τ^-, τ^+ and $\tau \in [\tau^-, \tau^+]$ it holds

$$\varphi(\tau) \leq \max\{\varphi(\tau^-), \varphi(\tau^+)\} . \quad (\text{A.7})$$

Therefore,

$$\begin{aligned}
g(\boldsymbol{\theta}^*) &\leq \max \left\{ (1 + c_{f,h}) \int_{-1}^1 K(t) f(x_0) (\log f(x_0) + \log(1 + c_{f,h}) - 1) dt, \right. \\
&\quad \left. (1 - c_{f,h}) \int_{-1}^1 K(t) f(x_0) (\log f(x_0) + \log(1 - c_{f,h}) - 1) dt \right\} \quad (\text{A.8}) \\
&= f_0 I_K(1 \pm c_{f,h}) \left((\log f_0 - 1) + \log(1 \pm c_{f,h}) \right) ,
\end{aligned}$$

$$g(\boldsymbol{\theta}^*) - g(\boldsymbol{\theta}^\circ) \leq f_0 I_K \left((1 \pm 1) c_{f,h} \log f_0 \mp c_{f,h} + (1 \pm c_{f,h}) \log(1 \pm c_{f,h}) \right) \quad (\text{A.9})$$

We see that the difference between $g(\boldsymbol{\theta}^\circ)$ and $g(\boldsymbol{\theta}^*)$ is small because $\log(1 \pm c_{f,h})$ is of order $c_{f,h}$, which is of order $O(h)$, $h \rightarrow 0$.

From the Taylor expansion, we have for some $\bar{\boldsymbol{\theta}}$:

$$g(\boldsymbol{\theta}^\circ) = g(\boldsymbol{\theta}^*) + \nabla g(\boldsymbol{\theta}^*)^\top (\boldsymbol{\theta}^\circ - \boldsymbol{\theta}^*) - \frac{1}{2} \|d_0^2(\bar{\boldsymbol{\theta}})(\boldsymbol{\theta}^\circ - \boldsymbol{\theta}^*)\|^2 , \quad (\text{A.10})$$

$$\begin{aligned}
\|d_0(\bar{\boldsymbol{\theta}})(\boldsymbol{\theta}^\circ - \boldsymbol{\theta}^*)\|^2 &= 2(g(\boldsymbol{\theta}^*) - g(\boldsymbol{\theta}^\circ)) \\
&\leq 2f_0 I_K \left((1 \pm 1) c_{f,h} \log f_0 \mp c_{f,h} + (1 \pm c_{f,h}) \log(1 \pm c_{f,h}) \right) \\
&= f_0 \cdot \phi_1^2 . \quad (\text{A.11})
\end{aligned}$$

Now we are going to perform the trick, which has a bit of asymptotical and implicit flavour. Let $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2 \in [\boldsymbol{\theta}^\circ, \boldsymbol{\theta}^*]$ — two points on the segment with the ends $\boldsymbol{\theta}^\circ$ and $\boldsymbol{\theta}^*$.

$$\varepsilon^2 = \sup_{\boldsymbol{\theta}_1, \boldsymbol{\theta}_2} \|\mathbf{I}_p - D^{-1}(\boldsymbol{\theta}_1) D^2(\boldsymbol{\theta}_2) D^{-1}(\boldsymbol{\theta}_1)\| \quad (\text{A.12})$$

Note that the following chain of inequalities is satisfied:

$$\begin{aligned}
\varepsilon^2 &\leq \left(\sup_{t \in [-1, 1]} \exp(\boldsymbol{\Psi}^\top(t) d_0^{-1} d_0(\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2)) - 1 \right)^2 \\
&\lesssim \sup_{t \in [-1, 1]} \boldsymbol{\Psi}(t)^\top d_0^{-2} \boldsymbol{\Psi}(t) \cdot (\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2)^\top d_0^2(\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2) \\
&\leq (1 + \varepsilon) \sup_{t \in [-1, 1]} \boldsymbol{\Psi}(t)^\top d_0^{-2}(\boldsymbol{\theta}^\circ) \boldsymbol{\Psi}(t) \cdot (1 + \varepsilon) (\boldsymbol{\theta}^\circ - \boldsymbol{\theta}^*)^\top d_0^2(\bar{\boldsymbol{\theta}})(\boldsymbol{\theta}^\circ - \boldsymbol{\theta}^*) \\
&\leq (1 + \varepsilon)^2 \mathbf{c}_1^2 \phi_1^2 .
\end{aligned}$$

The terms f_0 and f_0^{-1} cancelled out, because they appear both in $\|d_0(\bar{\boldsymbol{\theta}})(\boldsymbol{\theta}^\circ - \boldsymbol{\theta}^*)\|$ and $\sup_t \boldsymbol{\Psi}(t)^\top d_0^{-2}(\boldsymbol{\theta}^\circ) \boldsymbol{\Psi}(t) = f_0^{-1} \mathbf{c}_1^2$. Therefore,

$$\varepsilon \leq (1 + \varepsilon) \mathbf{c}_1 \phi_1, \quad \varepsilon \leq \frac{\mathbf{c}_1 \phi_1}{1 - \mathbf{c}_1 \phi_1} = O(h^{1/2}) . \quad (\text{A.13})$$

We can also notice that for any vector \mathbf{v} it holds

$$\mathbf{v}^\top d_0(\boldsymbol{\theta}^\circ) \mathbf{v} \leq (1 + \varepsilon) \mathbf{v}^\top d_0(\bar{\boldsymbol{\theta}}) \mathbf{v} , \quad (\text{A.14})$$

therefore

$$\|d_0(\boldsymbol{\theta}^\circ)(\boldsymbol{\theta}^\circ - \boldsymbol{\theta}^*)\|^2 \leq (1 + \varepsilon)f_0\phi_1^2 \lesssim f_0\phi_1^2(1 - \mathbf{c}_1\phi_1)^{-1} . \quad (\text{A.15})$$

End of the proof of lemma 5.

Remark 9. We have used the symbol “ \lesssim ” and an approximation “ $\exp(t) - 1 \lesssim t$ ”. This means that if t is close to zero, then the function $\exp(t) - 1$ is bounded by some linear function $\text{lin}(t) = \text{coeff} \cdot t$. It is enough to guarantee that $\boldsymbol{\theta}^\circ - \boldsymbol{\theta}^*$ is small when $h \rightarrow 0$. It is true because the matrix $d_0^2(\boldsymbol{\theta}^\circ)$ is positive-definite and continuous at the vicinity of $\boldsymbol{\theta}^\circ$ and the difference $g(\boldsymbol{\theta}^\circ) - g(\boldsymbol{\theta}^*)$ is close to zero.

After describing the closeness of $D_n^2(\boldsymbol{\theta}^\circ), D_n^2(\boldsymbol{\theta}^*)$ we describe the closeness of $D_n^2(\boldsymbol{\theta}^\circ)$ and $D_n^2(\boldsymbol{\theta}^*)$, and $\boldsymbol{\theta}^* \approx \boldsymbol{\theta}^\circ$ in metric generated by curvature matrix $d_0^2(\boldsymbol{\theta}^\circ)$.

Lemma 6. Let $I_K = \int_{-1}^1 K(t)dt \leq 2^d$. For vectors $\boldsymbol{\theta}^\circ, \boldsymbol{\theta}^*$ the following holds:

$$\|\mathbf{I}_p - D(\boldsymbol{\theta}^*)D^{-2}(\boldsymbol{\theta}^\circ)D(\boldsymbol{\theta}^*)\| \leq (1 + c_{f,h})B_{p,h} - 1 , \quad (\text{A.16})$$

$$\|d_0(\boldsymbol{\theta}^\circ)(\boldsymbol{\theta}^\circ - \boldsymbol{\theta}^*)\|^2 \leq I_K(f(x_0))^3(c_{f,h} - \log B_{p,h})^2 . \quad (\text{A.17})$$

Proof. Denote $\varphi(x) = \log f(x)$, $t = (x - x_0)/h$. According to the lemma 8, the quantity can be bounded by

$$\begin{aligned} \exp\left(\boldsymbol{\Psi}^\top(t)(\boldsymbol{\theta}^* - \boldsymbol{\theta}^\circ)\right) - 1 &= \exp\left((x - x_0)\varphi'(x_0) + \frac{(x - x_0)^2}{2!}\varphi''(x_0) + \right. \\ &\quad \left. \dots + \frac{(x - x_0)^{p-1}}{(p-1)!}\varphi^{(p-1)}(x_0)\right) - 1 \\ &\leq \exp(\log f(x) - \log f(x_0) - \log B_{p,h}) - 1 \\ &\leq (1 + c_{f,h})B_{p,h}^{-1} - 1 . \end{aligned}$$

Next,

$$\begin{aligned} \|d_0(\boldsymbol{\theta}^\circ)(\boldsymbol{\theta}^* - \boldsymbol{\theta}^\circ)\|^2 &= f(x_0) \int_{-1}^1 K(t) \left[(\boldsymbol{\theta}^* - \boldsymbol{\theta}^\circ)^\top \boldsymbol{\Psi}(t) \right] \left[(\boldsymbol{\theta}^* - \boldsymbol{\theta}^\circ)^\top \boldsymbol{\Psi}(t) \right]^\top dt \\ &= f(x_0) \int_{-1}^1 K(t) \left[f(x) - f(x_0) - \log B_{p,h} \right]^2 dt \\ &\leq I_K f(x_0)^3 (c_{f,h} - \log B_{p,h})^2 . \end{aligned} \quad (\text{A.18})$$

End of the proof of lemma 6.

Remark 10. Since $c_{f,h} = O(h)$, $B_{p,h} = 1 + O(h^p)$ the right hand side is of order $O(h)$.

Lemma 7. Let $\text{pr}_2(x_0) = \int_{-1}^1 K((x - x_0)/h)f^2(x)dx$, $g(\boldsymbol{\theta}) = (nh^d)^{-1}\mathbf{E}\mathbf{L}(\boldsymbol{\theta})$. Then the following inequality holds:

$$\|d_0^{-1}(\boldsymbol{\theta}^\circ)(\nabla g(\boldsymbol{\theta}^*) - \nabla g(\boldsymbol{\theta}^*))\|^2 \leq p \cdot (1 - B_{p,h})^2 h^{-d} \text{pr}_2(x_0) . \quad (\text{A.19})$$

Proof. Let $x = x_0 + t \cdot h$. Since $\nabla g(\boldsymbol{\theta}^*) = 0$, the difference between gradients is equal to

$$\nabla g(\boldsymbol{\theta}^\bullet) = \int_{\mathcal{X}} K(t) \boldsymbol{\Psi}(t) [f(x) - \exp(\boldsymbol{\Psi}^\top \boldsymbol{\theta}^\bullet)] dt . \quad (\text{A.20})$$

Denote $\delta(t) = \sqrt{K(t)}(f(x) - \exp(\boldsymbol{\Psi}^\top(t) \boldsymbol{\theta}^\bullet))$, and $\boldsymbol{\psi}(t) = \sqrt{K(t)} \boldsymbol{\Psi}(t)$. After applying lemma 9 (analog of the Cauchy-Schwarz inequality for vectors and matrices), we obtain:

$$\int_{-1}^1 \boldsymbol{\psi}(t) \delta(t) dt \int_{-1}^1 \boldsymbol{\psi}^\top(t) \delta(t) dt \leq \int_{-1}^1 \boldsymbol{\psi}(t) \boldsymbol{\psi}^\top(t) dt \int_{-1}^1 \delta^2(t) dt . \quad (\text{A.21})$$

Thus,

$$\nabla g(\boldsymbol{\theta}^\bullet) \nabla g(\boldsymbol{\theta}^\bullet)^\top \leq d_0^2(\boldsymbol{\theta}^\circ) \int_{-1}^1 \delta^2(t) dt . \quad (\text{A.22})$$

This allows to finish the proof:

$$\|d_0^{-1}(\boldsymbol{\theta}^\circ) \nabla g(\boldsymbol{\theta}^\bullet)\|^2 = \text{Tr } \nabla g(\boldsymbol{\theta}^\bullet)^\top d_0^{-2}(\boldsymbol{\theta}^\circ) \nabla g(\boldsymbol{\theta}^\bullet) \quad (\text{A.23})$$

$$= \text{Tr } \nabla g(\boldsymbol{\theta}^\bullet) \nabla g(\boldsymbol{\theta}^\bullet)^\top d_0^{-2}(\boldsymbol{\theta}^\circ) \quad (\text{A.24})$$

$$\leq p \int_{-1}^1 \delta^2(t) dt . \quad (\text{A.25})$$

The integral is bounded using the bias definition from section 3.2:

$$\int_{-1}^1 \delta^2(t) dt = \int_{-1}^1 K(t) [f(x) - \exp(\boldsymbol{\Psi}^\top \boldsymbol{\theta}^\bullet)]^2 dt \quad (\text{A.26})$$

$$\leq \int_{-1}^1 K(t) f(x)^2 [1 - B_{p,h}]^2 dt \quad (\text{A.27})$$

$$= (1 - B_{p,h})^2 \cdot h^{-d} \text{pr}_2(x_0) . \quad (\text{A.28})$$

End of the proof of lemma 7.

A.2 Facts from Linear Algebra

Lemma 8. If matrices $A, B \in \mathbb{R}^{p \times p}$ have the form

$$A^2 = \int_{\mathcal{X}} \boldsymbol{\Psi}(x) \lambda_A(x) \boldsymbol{\Psi}(x)^\top dx , \quad B^2 = \int_{\mathcal{X}} \boldsymbol{\Psi}(x) \lambda_B(x) \boldsymbol{\Psi}(x)^\top dx , \quad (\text{A.29})$$

and $\lambda_A(x), \lambda_B(x) > 0$, then the eigenvalue set of the quotient $B^{-1} A^2 B^{-1}$ belongs to the interval

$$\left[\min_{x \in \mathcal{X}} \frac{\lambda_A(x)}{\lambda_B(x)}, \max_{x \in \mathcal{X}} \frac{\lambda_A(x)}{\lambda_B(x)} \right] . \quad (\text{A.30})$$

Proof. Let us introduce self-adjoint operators $\Lambda_A, \Lambda_B: L_2[\mathcal{X}] \rightarrow L_2[\mathcal{X}]$, $\Psi: \mathbb{R}^p \rightarrow L_2[\mathcal{X}]$:

$$\Lambda_A f(x) = \lambda_A(x) f(x) \ , \quad \Lambda_B f(x) = \lambda_B(x) f(x) \ , \quad \Psi \mathbf{v} = \Psi(\mathbf{x})^\top \mathbf{v} \ . \quad (\text{A.31})$$

The scalar product takes form

$$\langle g(x), \Lambda f(x) \rangle = \int_{\mathcal{X}} g(x) \lambda(x) f(x) dx \ . \quad (\text{A.32})$$

Then matrices A^2, B^2 can be rewritten in the form

$$A^2 = \Psi^* \Lambda_A \Psi \ , \quad B^2 = \Psi^* \Lambda_B \Psi \ . \quad (\text{A.33})$$

It is not difficult to check that $\Lambda_B \succeq \min_{x \in \mathcal{X}} \frac{\lambda_B(x)}{\lambda_A(x)} \Lambda_A$. Therefore, for operator Ψ it holds

$$\Psi^* \Lambda_B \Psi \succeq \min_{x \in \mathcal{X}} \frac{\lambda_B(x)}{\lambda_A(x)} \Psi^* \Lambda_A \Psi \ , \quad \lambda_{\max}(B^{-1} A^2 B^{-1}) \geq \max_{x \in \mathcal{X}} \frac{\lambda_A(x)}{\lambda_B(x)} \ . \quad (\text{A.34})$$

Similar argument is suitable for the lower bound.

End of the proof of lemma 8.

Lemma 9. *Let $\psi(t): [-1, 1] \rightarrow \mathbb{R}^p$ be some vector-valued integrable function, $\delta(t): [-1, 1] \rightarrow \mathbb{R}$ — integrable function. Then the following matrix inequality holds:*

$$\int_{-1}^1 \psi(t) \delta(t) dt \int_{-1}^1 \psi^\top(t) \delta(t) dt \leq \int_{-1}^1 \psi(t) \psi^\top(t) dt \int_{-1}^1 \delta^2(t) dt \quad (\text{A.35})$$

Proof. Consider the matrix-valued non-negative integral:

$$I = \int_{-1}^1 d\tau \int_{-1}^1 dt [\psi(\tau) \delta(t) - \psi(t) \delta(\tau)] [\psi(\tau) \delta(t) - \psi(t) \delta(\tau)]^\top \geq 0 \quad (\text{A.36})$$

Denote $\int_{-1}^1 \psi(t) \delta(t) dt = A$, $\int_{-1}^1 \psi(t) \psi^\top(t) dt = M$, $\int_{-1}^1 \delta^2(t) dt = D$.

The integral I can be expanded and rewritten as:

$$I = \int_{-1}^1 \int_{-1}^1 \left[\delta(t)^2 \psi(\tau) \psi^\top(\tau) + \delta(\tau)^2 \psi(t) \psi^\top(t) \right. \quad (\text{A.37})$$

$$\left. - \delta(t) \delta(\tau) \psi(t) \psi^\top(\tau) - \delta(t) \delta(\tau) \psi(\tau) \psi^\top(t) \right] \quad (\text{A.38})$$

$$= 2AA^\top - 2DM \quad (\text{A.39})$$

Therefore,

$$AA^\top \leq DM \quad (\text{A.40})$$

End of the proof of lemma 9.

Lemma 10. Let $\Psi(t) = (1, t, t^2, \dots, t^{p-1})^\top$. Consider the matrix

$$A^2 = \int_{-1}^1 \Psi(t) \Psi^\top(t) dt . \quad (\text{A.41})$$

Then the polynomial defined by

$$P(t) = \Psi^\top(t) A^{-2} \Psi(t) \quad (\text{A.42})$$

attains its maximal value at points $t = \pm 1$, and this value equals to $p^2/2$. Moreover, the fact is still valid if we consider $\Psi(t) = (1, P_1(t), P_2(t), \dots, P_{p-1}(t))^\top$, where the polynomials $P_i(t)$ have degrees less than p and form a basis in the space of polynomials with degree less than p .

Remark 11. We formulated the following lemma experimentally and the proof for our guess was kindly presented by Ilya Bogdanov [13] at Mathoverflow. Some interesting properties of the polynomial $P(t)$ are listed in the discussion. The claim can be probably generalized to higher-dimensional case, but we still don't know whether it is possible to treat non-uniform kernel case efficiently. The shape of the polynomial $P(t)$ suggests that if we “suppress” its behaviour at the tails, say by choosing appropriate kernel function, this constant can be reduced significantly.

Proof. It is well-known that Legendre polynomials form the orthogonal basis for the space of polynomials, defined on the segment $[-1, 1]$ with respect to the scalar product

$$\langle f, g \rangle = \int_{-1}^1 f(t) g(t) dt . \quad (\text{A.43})$$

The ordinary Legendre polynomials $L_k(t)$ are equal to ± 1 at the ends of the interval $[-1, 1]$, and their scalar product equals to

$$\langle L_i(t), L_j(t) \rangle = \delta_{ij} \cdot \frac{2}{2j+1} . \quad (\text{A.44})$$

We consider their normed versions $\tilde{L}_j(t) = \sqrt{\frac{2j+1}{2}} L_j(t)$ so that the basis $\mathbf{L} = (L_0, L_1, \dots, L_{p-1})$ is orthonormal, i.e. $\int_{-1}^1 \mathbf{L}(t) \mathbf{L}^\top(t) dt = \mathbf{I}_p$.

Since the polynomials $1, t, t^2, \dots, t^{p-1}$ have degrees less than p and are linearly independent, the basis $\Psi(t)$ can be transferred into the Legendre polynomial basis $\mathbf{L}(t)$ with some transition matrix S : $\Psi = S\mathbf{L}$. Substituting this value into the expression (A.42), we obtain:

$$P(t) = \mathbf{L}^\top(t) S^\top \left(\int_{-1}^1 S \mathbf{L}(t) \mathbf{L}^\top(t) S^\top dt \right)^{-1} S \mathbf{L}(t) \quad (\text{A.45})$$

$$= \mathbf{L}^\top(t) S^\top S^{-\top} \mathbf{I}_p S^{-1} S \mathbf{L}(t) \quad (\text{A.46})$$

$$= \mathbf{L}^\top(t) \mathbf{L}(t) = \sum_{j=0}^{p-1} \tilde{L}_j^2(t) . \quad (\text{A.47})$$

It is well-known that Legendre polynomials $L_j(t)$ are uniformly bounded $|L_j(t)| \leq 1$, and the maximum is attained at $t = \pm 1$. Therefore, the maximal value of the polynomial $P(t)$ on the segment $[-1, 1]$ equals to

$$P(\pm 1) = \frac{1}{2} \sum_{j=0}^{p-1} (2j+1) = \frac{p^2}{2} . \quad (\text{A.48})$$

End of the proof of lemma 10.

Lemma 11. Let $\text{pr}_1(x_0) = \int_{-1}^1 K((x - x_0)/h) f(x) dx$,

$$C_{V,f}^2 = nh^d \sup_{t \in [-1, 1]} K(t)^2 \boldsymbol{\Psi}^\top(t) V_n^{-2}(f(x)) \boldsymbol{\Psi}(t) . \quad (\text{A.49})$$

Then we have an exact relationship:

$$C_{V,f}^2 = \sup_{t \in [-1, 1]} K(t)^2 \boldsymbol{\Psi}(t)^\top \left[\int K f \boldsymbol{\Psi} \boldsymbol{\Psi}^\top d\tau \right]^{-1} \boldsymbol{\Psi}(t) + \frac{h^d K(t)^2}{1 - \text{pr}_1(x_0)} , \quad (\text{A.50})$$

which leads to the inequality

$$C_{V,f}^2 \leq (1 - c_{f,h})^{-1} f(x_0)^{-1} \mathbf{c}_2^2 + \frac{h^d}{1 - \text{pr}_1(x_0)} . \quad (\text{A.51})$$

Proof. By the definition of $V_n(f(x))$, we can express

$$n^{-1} V_n^2(f(x)) = \int_{-1}^1 K f \boldsymbol{\Psi} \boldsymbol{\Psi}^\top dx - \int_{-1}^1 K f \boldsymbol{\Psi} dx \int_{-1}^1 K f \boldsymbol{\Psi}^\top dx , \quad (\text{A.52})$$

and replacing dx with $h^d dt$, we obtain

$$(nh^d)^{-1} V_n^2(f(x)) = \int_{-1}^1 K f \boldsymbol{\Psi} \boldsymbol{\Psi}^\top dt - h^d \int_{-1}^1 K f \boldsymbol{\Psi} dt \int_{-1}^1 K f \boldsymbol{\Psi}^\top dt , \quad (\text{A.53})$$

which can be denoted as $A - h^d \mathbf{u} \mathbf{u}^\top$ with $A = \int K f \boldsymbol{\Psi} \boldsymbol{\Psi}^\top dt$, $\mathbf{u} = \int K f \boldsymbol{\Psi} dt$. Then we apply Sherman–Morrison formula for one-rank inverse matrix updates:

$$(A - \lambda \mathbf{u} \mathbf{u}^\top)^{-1} = A^{-1} + \lambda \frac{A^{-1} \mathbf{u} \mathbf{u}^\top A^{-1}}{1 - \lambda \mathbf{u}^\top A^{-1} \mathbf{u}} . \quad (\text{A.54})$$

Since the basis $\boldsymbol{\Psi}$ has important property that first element of the basis $\psi_0(t)$ is constant 1, we note that

$$\mathbf{u} = A \cdot [1, 0, \dots, 0]^\top, \quad A^{-1} \mathbf{u} = [1, 0, \dots, 0] , \quad (\text{A.55})$$

and this allows to simplify the above expression. Multiplying by $\boldsymbol{\Psi}$ from the right and from the left, we obtain final exact expression:

$$C_{V,f}^2 = \sup_{t \in [-1, 1]} \boldsymbol{\Psi}^\top A^{-1} \boldsymbol{\Psi} + \frac{\lambda}{1 - \lambda \int K(t) f(x_0 + ht) dt} , \quad (\text{A.56})$$

where $\lambda = h^d$. The consequent inequality is straightforward.

End of the proof of lemma 11.

A.3 Deviation Bounds for Quadratic Forms

Lemma 12 (Laurent and Massart, [15]). *Let $A \in \mathbb{R}^{p \times p}$, $\boldsymbol{\xi} \in \mathbb{R}^p$ be a random vector of the form $\boldsymbol{\xi} = A^{-1}\mathbf{u}$, where $\mathbf{E}[\mathbf{u}] = 0$. Denote $V^2 = \text{Var } \mathbf{u}$, $\mathfrak{a}^2 = \lambda_{\max}(A^{-1}VA^{-1})$.*

Suppose that the vector $V^{-1}\mathbf{u}$ satisfies the condition

$$\log \mathbf{E} \exp(\boldsymbol{\gamma}^\top V^{-1}\mathbf{u}) \leq \frac{\nu_0^2 \|\boldsymbol{\gamma}\|^2}{2}, \quad \boldsymbol{\gamma} \in \mathbb{R}^p. \quad (\text{A.57})$$

Then for each $z > 0$, $z \leq \mathfrak{g}^2/4$

$$\mathbf{P}(\|\boldsymbol{\xi}\| > \zeta(p, z)) \leq 2e^{-z} + 8.4e^{-\mathfrak{g}^2/4}, \quad (\text{A.58})$$

where $\zeta(p, z)$ is defined by

$$\zeta(p, x) = \mathfrak{a}\nu_0(\sqrt{p} + \sqrt{2z}). \quad (\text{A.59})$$

Acknowledgments. First of all, I would like to thank Vladimir Spokoiny for providing the formulation of the problem considered in this article. I would also like to thank Fedor Goncharov, Ekaterina Krymova, Alexey Balitsky, Alexander Cigler and Nazar Buzun for fruitful discussions and providing helpful comments. I thank Elena Chernousova for pointing out some errors and for reviewing this paper, because her remarks helped me to improve it. I am grateful to Ilya Bogdanov [13], who pointed me the proof of lemma 10 at mathoverflow.net and to the whole this community.

References

1. Spokoiny, V.: Bernstein - von Mises Theorem for growing parameter dimension. (2013) <http://arxiv.org/abs/1302.3430>
2. Spokoiny, V.: Parametric estimation. Finite sample theory. Ann. Statist. Volume 40, Number 6, 2877–2909 (2012). <http://arxiv.org/abs/1111.3029>
3. Loader, C.: Local Likelihood Density Estimation. Ann. Statist. Volume 24, Number 4, 1602–1618 (1996)
4. Tsybakov, Alexandre B.: Introduction to Nonparametric Estimation. 1st ed. New York: Springer (2008)
5. Silverman, B.W.: Density Estimators for Statistics and Data Analysis. Monographs on Statistics and Applied Probability, London: Chapman and Hall (1986)
6. Nadaraya, E. A.: On Estimating Regression. Theory of Probability and its Applications. Volume 9, Number 1, 141–142 (1964)
7. Watson, G. S.: Smooth regression analysis. The Indian Journal of Statistics, Series A. Volume 26, Number 4, 359–372 (1964)
8. Tibshirani, R., Hastie, T.: Local Likelihood Estimation. J. Amer. Statist. Assoc. 82, 559–5567 (1987)
9. Talagrand, M.: Upper and Lower Bounds for Stochastic Processes. A Series of Modern Surveys in Mathematics, Springer Berlin Heidelberg (2014)
10. Jones, M. C.: Variable kernel density estimates and variable kernel density estimates. Austral. J. Statist. 32, 361–371 (1990)

11. Lepski, O. and Spokoiny, V.: Optimal pointwise adaptive methods in nonparametric estimation. *Ann. Statist.* Volume 25, Number 6, 2512–2546 (1997)
12. Spokoiny V., Zhilova M.: Bootstrap confidence sets under a model misspecification. Berlin, WIAS Preprint no. 1992 (2015). <http://arxiv.org/abs/1410.0347>
13. <http://mathoverflow.net/questions/236547/why-polynomial-psi-topt-a-1-psit-attains-maximum-on-1-1-at-t/236567>
14. Zhilova M.: Simultaneous likelihood-based bootstrap confidence sets for a large number of models. SFB 649 Discussion Paper 2015-031, <http://arxiv.org/abs/1506.05779>
15. Laurent, B. and Massart, P.: Adaptive estimation of a quadratic functional by model selection. *Ann. Statist.*, 28(5), 1302–1338 (2000)